Affect-Based Indexing and Retrieval of Films

Ching Hau Chan
School of Computing &
Centre for Digital Video Processing
Dublin City University
Glasnevin, Dublin 9, Ireland
chchan@computing.dcu.ie

Gareth J. F. Jones
School of Computing &
Centre for Digital Video Processing
Dublin City University
Glasnevin, Dublin 9, Ireland
gjones@computing.dcu.ie

ABSTRACT

Digital multimedia systems are creating many new opportunities for rapid access to content archives. In order to explore these collections using search applications, the content must be annotated with significant features. An important and often overlooked aspect of human interpretation of multimedia data is the affective dimension. Affective labels of content can be extracted automatically from within multimedia data streams. These can then be used for content-based retrieval and browsing. In this study affective features extracted from multimedia audio content are mapped onto a set of keywords with predetermined emotional interpretations. These labels are then used to demonstrate affect-based retrieval on a range of feature films.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

affect-based retrieval, content-based retrieval, affective labelling of multimedia

1. INTRODUCTION

Multimedia systems now enable enormous volumes of high quality content to be distributed to users on demand. In order to most effectively exploit this data, users require tools for selection and classification of relevant content. A core technology for such tools is the automated analysis of the content to index significant features. The content of multimedia data relates both to clearly defined semantic units such as words and visual objects, and to more subjective interpretation, such as the emotional component represented in the content. Existing work on multimedia content analysis has concentrated largely on recognition of the objective features. A number of current large-scale prototype multimedia digital library systems based on using these features have been developed, for example [1][2]. The feature indexing tools within these systems rely on extraction of low-level feature analysis such as the pitch or frequency spectrum of the audio or the colour distribution and texture in the video. These low-level features can also be used to recognize non-verbal features in multimedia audio such as special effects [3], and can be interpreted in terms of their contribution to the affective labelling of multimedia data [4] [5]. In this paper we describe initial work on a system for the analysis, interpretation and retrieval of films based on affective features.

The affective dimension is an important natural component of human classification and retrieval of information. For example, music and sound effects are often used to great effect to convey emotional information in films and commercials. Recognizing multimedia content contributing to this dimension and using it to automatically label the significant affective features potentially allows a new modality for user interaction with multimedia content. For example, to search for sad or happy sections of a film, or compare the structure of films in terms of their emotional development. Such methods might also be used to label emotional expressions in recordings of discussions or debates, e.g. anger or sadness, alongside the verbal transcriptions currently used to manage this content. Thus a different response may be produced for the same transcribed content depending on their affective context. The system presented in this paper is a step toward labelling the emotional dimension of multimedia data using low-level feature analysis, and their association with a predefined set of emotional labels. The affective metadata associated with multimedia documents labelled is then used as a means of document searching within an information retrieval (IR) system. A similar annotation idea is explored for manually assigned metadata in [6], which examines the structure of films based on the emotional priming of words appearing in "audio descriptions" provided for sight impaired viewers.

This paper is organized as follows: Section 2 reviews existing work in affect extraction and labelling, and outlines our affect annotation system, Section 3 describes the IR method used in our retrieval system, Section 4 gives examples from our current experimental investigations, and Section 5 summarizes our conclusions and outlines our ongoing work.

MM'05, November 6–11, 2005, Singapore.
.

2. REPRESENTATION, EXTRACTION AND LABELLING OF AFFECTIVE STATES

Our approach to affect annotation combines work from fundamental representation of emotional states, with methods to recognize these states in audio-visual content.

2.1 Dimensional Representation of Affective States

It is demonstrated in [7][8] that human affective states can be described in terms of three basic dimensions: *arousal*, *valence* and *control* (dominance). Arousal is a continuous response ranging from one extreme of sleep through intermediate states of drowsiness and then alertness through to frenzied excitement. Valence describes the degree of pleasantness-unpleasantness ranging from a "positive" response associated with extreme happiness or ecstasy through to a "negative" response resulting from extreme pain or unhappiness. Control can be useful to distinguish between affective states with similar arousal and valence, with differing feelings of control or influence over events ranging from total control to no control. From this definition "emotion" not only describes occasional passionate incidents, but rather a person is viewed as being in an emotional state within this three dimensional space at all times.

It has been observed that the control dimension plays only a limited role in extracting affective content from multimedia data [9]. In our work we thus follow the strategy adopted in [4] of measuring only arousal and valence values. Measuring arousal and valence levels across a multimedia document enables us to extract information of its affective content. A classic way to combine measurements of arousal and valence is the *affect curve*. This plots arousal on the y-axis against valence on the x-axis. Each point of this curve corresponds to an affective state.

2.2 Extraction of Arousal and Valence from Multimedia Data

Methods for extracting arousal and valence information of the speech signal in audio-visual recordings are introduced in [4], where the techniques are based on a combination of analysis of both audio and visual features. Our current system ultilizes only audio feature extraction, but focuses on the analysis of all the audio data, including speech, music, special effects and silence. The following subsections outline our methods to extract arousal and valence information.

2.2.1 Arousal Modelling

The arousal level is modelled based on the energy of the audio signal. The sampled audio data is first divided into overlapping frames. If it is assumed that the audio signal changes relatively slowly within a short interval and taking a suitably small frame, the energy is calculated for each frame [3]. In our experiments using a frame size of 20ms and a sampling frequency of 44.1kHz with a 2/3 overlap of frames was found to be effective. The energy level can vary significantly between adjacent frames. However, variations in emotion do not occur abruptly in this manner. For example, excitement may increase as the action develops in a scene, and then decrease gradually afterwards. In order to more appropriately model overall changes in arousal levels as the affective state of the content changes, the measured short-time energy is convoluted with a long Kaiser window as described in [4].

2.2.2 Valence Modelling

It is noted in [11] that extraction of features to determine valence values is much more difficult than arousal. In our system we currently use pitch to measure the magnitude and sign of valence, as suggested in [4]. The short-time fundamental frequency is calculated based on peak detection from the spectrum of the sound. This is done using an autoregressive model of order 40 as described in [3]. Short-time fundamental frequency values are calculated for the same 20ms frames as used to calculate the arousal values.

Valence varies from a "neutral feeling" towards either a negative or positive response. The system needs to identify a frequency corresponding to this neutral feeling; points whose pitch value varies from this can then be labelled with a degree of positive or negative valence. The neutral pitch will vary with the audio source type and even between different speakers. In our current system we ignore these variations, and adopt the simple approach of assuming that the mean pitch value of the signal corresponds to the neutral state, and subtract this from all the points in the data stream. This method assumes that sources will have similar amounts of positive and negative response, which obviously may not be the case for some sources. We are exploring methods to compute a more accurate source dependent neutral frequency.

The normalized measured pitch values are subject to large changes between adjacent frames, similar to the variation in the short-time power used to measure the arousal, and for the same reason as before, we again smooth the valence function using a long Kaiser window.

2.3 Verbal Labelling of Affective States

The interpretation of the arousal and valence values extracted using the methods described in the previous section, is much easier if they can be identified with verbal labels corresponding to associated affective states. We do this by adopting a method based on results of experiments described in [6][7]. In the study described in [7] a group of test subjects defined 151 emotion-denoting keywords in terms of the dimensions of arousal, valence and dominance. Mean values of the dimension values were then computed across all subjects rating each keyword. The mean values were then transformed linearly to a scale ranging from -1 to $+1$ with a neutral value of 0. For example, the keyword "bold" has arousal 0.61, valence 0.44 and dominance 0.66, and "nonchalant" has arousal -0.25, valence 0.07 and dominance 0.11. Using the results from [7], we can associate each measured point on an affect curve plotting arousal against valence with the nearest keyword. To fit our arousal and valence results onto the same scale as the emotion-denoting keyword the values are scaled across the document into the range -1 to $+1$. This once again assumes that the full range of affective states is suitable for all multimedia data sources, exploring methods of determining affective range within individual documents is the subject of ongoing work.

The accuracy of the assignment of keywords to individual frames is likely to be fairly unstable and from our observation these words are unlikely to be those associated with describing the affective dimensions of multimedia document. The final stage of our labelling method is thus to associate each assigned keyword with one of 22 broader emotion classes proposed in [10]. In [6] these 22 emotions are expanded to a set of 627 emotion keywords using the WordNet ontology. For example, the emotion class "fear" is associ-

ated with the keyword "alarmed". We found that 50 of these 627 keywords were present in the list of 151 words; we then manually mapped the remaining 101 words onto the emotion classes using a standard English dictionary to identify the nearest emotion for each word. In operation each point on our automatically extracted audio affect curve is labelled with the nearest of the 151 emotion-denoting keywords and then mapped to the corresponding emotion class.

2.4 Labelling Dominant Emotions

The output of the system is a textual label of the nearest affective state for each 20ms frame. Taken across even a short piece of multimedia data, the resulting set of label will include instances of many different emotions. Many of these will be isolated frame labels not reflecting dominant emotions within a section of the data. These labels may reflect peripheral affective events, or more often will probably be inappropriate labels resulting from a combination of inaccurate modelling of the complex content of audio signal and the effects of smoothing with the Kaiser window. For multimedia content management applications we need to interpret the significant emotional impact of the audio signal. To do this we wish to label the dominant emotions across a multimedia document. In order to do this we adapt a technique from the field of text summarization introduced by Luhn [12].

Luhn reasoned that significant words in a phrase are significantly related if they are separated by not more than five high frequency insignificant stop words. He proposed a cluster significance score (CSS) factor, $CSS = SW^2/TW$ where SW = the number of significant words in a cluster, and TW = the total number of words in a cluster. Sentences in a document can then be ranked based on the clusters they contain, and the highest scoring sentences selected as a document summary.

We adapted this method to score multimedia affect label clusters as follows. Significant labels for each emotion were found by considering frame-level labels for this current emotion as significant and all others as insignificant. Clusters of emotion words were then located in a sequence of frame labels. When the distance between two frame labels of the current emotion exceeded an empirically determined limited, a new cluster was started. A significance score was then assigned to each cluster segment for each emotion using Luhn's metric. The individual clusters were scored for each of the 22 emotions. Clusters scoring above an empirically determined threshold were assigned as the emotional labels for the section of data under consideration. Note that this method allows the assignment of overlapping significant emotions for sections of data, and that the granularity of the assignments can be controlled by varying the threshold.

3. INFORMATION RETRIEVAL METHODS

Our affect-based film retrieval system enables users to search for documents with a desired emotional content by using ad hoc IR methods. The system adopts the Okapi IR model [13], shown to be very effective in many comparative IR evaluation exercises in recent years.

The search terms for the IR system (the affect labels assigned to each document) are first weighted using the Okapi *combined weight* (cw), often known as BM25, [13]. The BM25 cw for a term is calculated as follows,

$$cw(i,j) = \frac{cfw(i) \times tf(i,j) \times (K1+1)}{K1 \times ((1-b) + (b \times ndl(j))) + tf(i,j)}$$

where $cw(i,j)$ represents the weight of term i in document j, $cfw(i) = \log(N/n(i))$ the standard collection frequency (inverse document frequency) weight, $n(i)$ is the total number of documents containing term i, and N is the total number of documents in the collection, $tf(i,j)$ is the document term frequency, and $ndl(j) = dl(j)$ Av. dl is the normalised document length where $dl(j)$ is the length of j. $K1$ and b are empirically selected tuning constants for a particular collection. The matching score for each document is computed by summing the weights of terms appearing in the query and the document, which are then returned to the user in order of decreasing matching score.

4. EXPERIMENTAL INVESTIGATION

In this section we present experimental results of our initial affect labelling and retrieval system for films. The first section illustrates the output of arousal and valence, how these are plotted onto an affect curve, and the textual annotation of emotions. The second section then describes examples from the use of our affect-based retrieval system.

4.1 Test Data

Our affect annotation system has so far been run on 4 feature films: *Ocean's Eleven*, *The Pelican Brief*, *Enigma*, and *The Road to Perdition*. For this preliminary study the films were divided into 10 minute sections which were processed and retrieved independently. For the purposes of our retrieval experiment, the 10 minute sections are treated as independent documents.

The affect label clustering parameters were set empirically based on observation of annotation output as: a maximum inter-frame cluster gap up to 5 for each emotion, and CSS threshold 0.5. Since the documents were of identical length in this study, the length normalization components of the BM25 term weighting have no effect. The BM25 $K1$ factor was set to 1.5 which is generally found to be useful [13]. These parameters will be explored more fully in a future study using a larger data set.

4.2 Affect Annotation

Figures 1 and 2 show example arousal and valence curves respectively for the third 10 minute section of the film *The Road to Perdition*. These are then plotted as an arousal curve in Figure 3. From this curve we can see the relationship between arousal and valence for each frame of the data. The points on the arousal curve are then labelled with the closest emotion as described in Section 2.3. Clusters for each emotion are then calculated using the procedure described in Section 2.4. Figure 4 shows the cluster output for six emotions for this clip. This shows a dominance of *fear* and *hate* in this section of the film. In this section of the film, the main character witnesses a murder, and and then fears for their own life.

4.3 Affect-Based Retrieval

To explore film retrieval using these labels we examined the response to four test search topics: (*anger*, *hate*, *resentment*); (*pity*, *gloat*); (*joy*, *love*); (*hate*, *joy*). These were selected to cover a range of similar and contrasting emotion

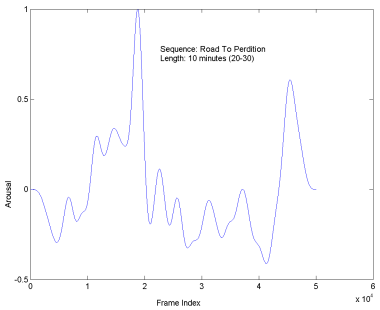

Figure 1: *Perdition*(3): Arousal.

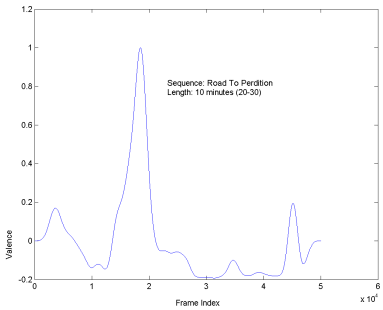

Figure 2: *Perdition*(3): Valence.

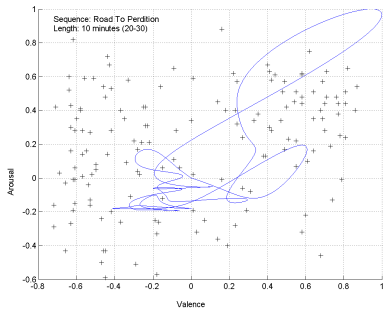

Figure 3: *Perdition* (3): Affect.

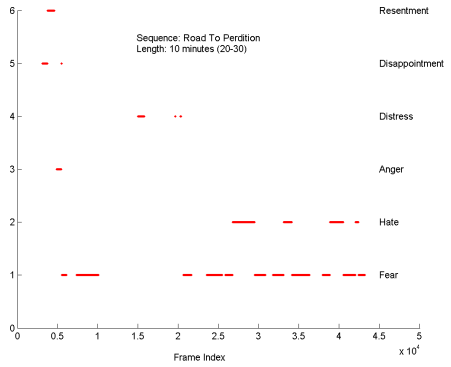

Figure 4: *Road to Perdition* (3): Emotions.

types. Examining the retrieved ranked lists we observed the following: playback of the highest ranked film clips for each topic generally indicated that they contained elements of the requested emotions; contiguous film sections were generally grouped together in the list, indicating that they contain similar emotional content; the contrasting emotions of the last topic produced a more varied list, indicating that these emotions are generally less well correlated in film scenes, as we would expect.

5. CONCLUSIONS AND FURTHER WORK

This paper has described our prototype system for affect-based indexing and retrieval from feature films. The results from our preliminary study are encouraging, illustrating that we are able to locate relevant scenes from within a small set of films. We are currently developing a much larger collection of affect annotated films as part of a formal IR test collection, including a set of search topics and manual relevance assessments. We are also extending our affect annotation scheme to incorporate visual features, such as those introduced in [4] which incorporate visual motion activity and density of shot cuts from the video stream as components in the arousal measure.

We also plan to extend our study to compare our affect annotation with those generated in [6] using manual audio descriptions. Results of this comparison may lead to methods of effectively combining the alternative annotation schemes to provide richer or more reliable affective labelling. We are also working on the incorporation of a temporal component into document and topic labels to enable us to query films with respect to their emotional profile.

6. REFERENCES

[1] Smeaton, A.F., Lee, H. and McDonald, K.: Experiences of Creating Four Video Library Collections with the Fischlar System, International Journal on Digital Libraries, 4(10) (2004) 42-44

[2] Hauptmann, A.G., Christel, M.G.: Successful Approaches in the TREC Video Retrieval Evaluations, Proceedings of ACM Multimedia 2004, New York City, ACM (2004) 668-675

[3] Zhang, T., Kuo, C. C. J.: Content-Based Audio Classification and Retrieval for Audiovisual Data Parsing, Kluwer Academic Publishers, (2001)

[4] Hanjalic, A., Xu., L.-Q.: Affective Video Content Representation and Modeling, IEEE Transactions on Multimedia, 7(1) (2005) 143-154

[5] Chan, C. H., and Jones, G. J. F.: Annotation of Multimedia Audio Data with Affective Labels for Information Management, Proceedings of PRIS 2005 - 5th International Workshop on Pattern Recognition in Information Systems, Miami, USA (2005) 94-103

[6] Salway, A. and Graham, M.: Extracting Information about Emotions in Films, Proceedings of ACM Multimedia, Berkeley, ACM (2003) 299-302

[7] Russell, J., Mehrabian, A.: Evidence for a Three-Factor Theory of Emotions, Journal of Research in Personality, 11 (1977) 273-294

[8] Bradley, M. M.: Emotional Memory: A Dimensional Analysis. In: van Groot, S., van de Poll, N.E., Sargeant, J. (eds.) The Emotions: Essays on Emotion Theory, Hillsdale, NJ: Erlbaum (1994) 97-134

[9] Dietz, R., Lang, A.: Affective Agents: Effects of Agent Affect on Arousal, Attention, Liking and Learning, Proceedings of the Third International Cognitive Technology Conference, San Francisco (1999)

[10] Ortony, A, Clore, G.L., Coolins, A.: The Cognitive Structure of Emotions, CUP (1988)

[11] Picard, R: Affective Computing, MIT Press (1997)

[12] Luhn, H. P.: The Automatic Creation of Literature Abstracts, IBM Journal of Research and Development, 2(2) (1958) 159-165

[13] S. E. Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., and M.Gatford, M.: Okapi at TREC-3. In D. K. Harman, editor, *Proceedings of TREC-3*, NIST (1995) 109-126